Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter

Nawal Ould Amer, Philippe Mulhem, Mathias Géry

Recherche de conversations dans les réseaux sociaux : modélisation et expérimentations sur Twitter.

Nawal Ould Amer[*,**] **— Philippe Mulhem** [*] **— Mathias Géry** [**]

[*] *Université de Grenoble Alpes, LIG, F-38000 Grenoble, France*
CNRS, LIG, F-38000 Grenoble, France
Nawal.Ould-Amer, Philippe.Mulhem@imag.fr
[**] *Université de Lyon, F-42023, Saint-Étienne, France,*
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
mathias.gery@univ-st-etienne.fr

RÉSUMÉ. *La problématique étudiée dans cet article est celle de l'indexation et de la recherche de conversations dans les réseaux sociaux. Une conversation est un ensemble de messages échangés entre utilisateurs, à la suite d'un message initial. La démarche proposée se base sur une modélisation probabiliste, et détaille en particulier l'utilisation d'informations sociales dans le réseau Twitter. Notre proposition est évaluée sur un corpus de conversations contenant plus de 50 000 tweets, et sur un ensemble de 15 requêtes tirées pour partie des campagnes TREC Microblog (Lin et Efron, 2013). Les résultats obtenus en combinant les éléments de contenu et les éléments sociaux sur ce corpus sont statistiquement significativement meilleurs que ceux de notre approche utilisant le contenu seul ainsi que ceux d'une approche à base de BM25.*

ABSTRACT. *The problem considered in this paper tackles the indexing and retrieval of conversations in social networks. A conversation is a set of messages exchanged between users, following an initial message. The proposed approach is based on a probabilistic modeling, focusing specifically on the use of social information available on the Twitter platform. Our proposal is evaluated on a corpus of conversations with more than 50,000 tweets, and a set of 15 queries drawn partly from TREC Microblog campaigns (Lin et Efron, 2013). The results obtained on this corpus are statistically significantly better than two content-only probabilistic approaches.*

MOTS-CLÉS : *Indexation de conversation, Modèle probabiliste, Collection de test.*

KEYWORDS: *Conversation indexing, Probabilistic modeling, Test collection.*

1. Introduction

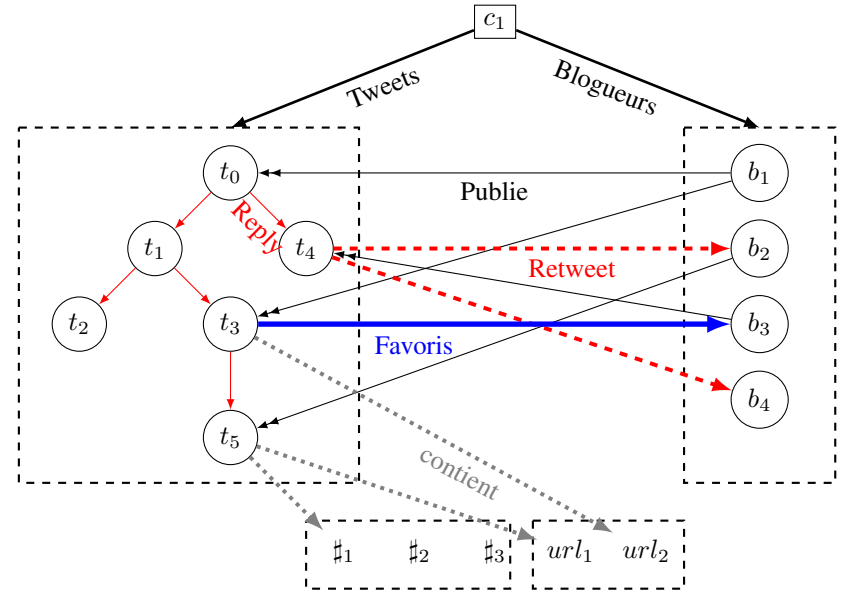
La recherche d'information sociale a pour objectif de retrouver des documents qui correspondent à un besoin d'information d'un utilisateur, tout en intégrant des éléments provenant de la participation des utilisateurs à des réseaux sociaux. Ces réseaux sociaux tel que Twitter offrent aux utilisateurs la possibilité de communiquer, d'interagir, et de répondre aux messages des autres en créant ainsi des conversations. Les travaux de recherche sur les microblogs se sont principalement intéressés à sélectionner séparément les tweets traitant un sujet donné. La totalité de la conversation est souvent négligée. Nous nous intéressons dans cet article à la recherche de *conversations* sociales, une conversation étant un ensemble de messages échangés par des utilisateurs sur un sujet donné. Notre idée est que qu'une conversation est potentiellement plus intéressante qu'un seul tweet comme source d'information pour une requête donnée.

Magnani définit une conversation comme étant un arbre où chaque nœud représente un message (tweet) posté par un utilisateur (blogueur) à un instant donné en réponse à un nœud parent (Magnani *et al.*, 2012). Semblablement à (Magnani *et al.*, 2012), nous définissons une conversation c dans le microblog Twitter comme un ensemble de tweets échangés entre les blogueurs b_j à l'aide de la fonction *Reply* de Twitter (cf. figure 1). Chaque tweet est publié à un instant donné en réponse à un tweet plus ancien (sauf pour le tweet initial). Dans la figure 1, le tweet t_4 est une réponse au tweet t_0. Les blogueurs sont liés par des relations d'abonnements. Un blogueur a la possibilité de publier un *tweet* à ses abonnées (*followers*), de rediffuser un *tweet* d'un autre blogueur (*"RT"*), en ajoutant éventuellement un commentaire, de répondre à un *tweet* (*"@username"*), de mentionner un blogueur dans un tweet (*@mention*), de mettre un *tweet* en favoris (*star*), d'ajouter dans le *tweet* des mots-clés appelés *hashtags* (*#motclé*) et de partager une ressource Web (*url*). La conversation c_1 contient donc un ensemble de tweets $t_0, t_1, ..., t_5$ qui sont publiés par les blogueurs $b_1, b_2, ..., b_4$. Le blogueur b_1 a publié le tweet t_0 qui est le tweet initial de la conversation c_1. Le blogueur b_2 répond *"Reply"* au tweet t_3 en publiant le tweet t_5 qui comporte le hashtag ($\#1$) et l'url (url_1). Les blogueurs b_2 et b_4 ont retweeté le tweet t_4 publié par le blogueur b_3 et le blogueur b_3 a mis en favoris le tweet t_3.

Une des difficultés de la recherche d'information sociale repose sur le fait qu'il existe une multitude de sources d'informations potentiellement utiles (comme le fait qu'un blogueur est populaire, qu'un tweet est considéré intéressant par de nombreux blogueurs, etc.) : la question est alors de déterminer lesquelles utiliser, comment les utiliser, et comment intégrer toutes ces sources pour obtenir une recherche d'information de qualité. Le travail décrit ici vise à intégrer, dans un modèle probabiliste clair, un certain nombre d'éléments identifiés et explicités, afin de réaliser un système de recherche d'information qui fournit des bons résultats sur des corpus de conversations.

Dans la suite de cet article, nous commençons par décrire en section 2 les travaux relatifs à notre problématique. La section 3 formalisera et présentera le modèle que nous proposons. La section 5 décrira les évaluations effectuées, précédée en section 4

Figure 1. *Exemple de conversation dans le microblog Twitter*



Retweet (rediffusion) : le tweet t_4 a été retweeté par le blogueur b_2.
Favoris : le tweet t_3 a été mis en favoris par le blogueur b_3.
Reply (réponse) : le tweet t_4 est une réponse (reply) au tweet t_0.
Publie : le blogueur b_1 a publié le tweet t_0.
Contient : le tweet t_5 contient un hashtag $\#_1$ et une url url_1.

par la description de la création de la collection de test sur laquelle nous avons mené nos expérimentations.

2. État de l'art

Différents travaux se sont intéressés à la recherche de conversations dans les structures sociales comme les blogs, les forums et les microblogs (Twitter). Nous classons ces travaux en deux catégories : la première basée sur le contenu textuel seul des conversations et la seconde basée sur le contenu textuel et les informations sociales.

La catégorie focalisée sur l'exploitation du contenu textuel se décompose en deux approches, suivant la manière de combiner les messages d'une conversation. La fusion précoce consiste à concaténer tous les messages d'une conversation dans un seul document global et la pertinence de la conversation est donnée par le score de correspondance entre la requête et ce document global. Dans cet axe, Seo, Elsas et Car-

bonell utilisent un modèle de langue sur le document global (Seo *et al.*, 2009 ; Elsas et Carbonell, 2009). La fusion tardive combine les scores de correspondance de chaque message de la conversation. Trois approches rentrent dans cet axe. La première prend en compte tous les messages, applique un modèle de langue sur chaque message et fusionne les scores pour obtenir un score global pour la conversation (Seo *et al.*, 2009 ; Elsas et Carbonell, 2009). La deuxième approche sélectionne les messages à utiliser avant de fusionner leur score. Cette sélection consiste à ne garder que les messages les plus pertinents lors du calcul de correspondance globale (Seo *et al.*, 2009 ; Elsas et Carbonell, 2009 ; Albaham et Salim, 2012). La troisième approche consiste à utiliser la structure des conversations. Le score global est alors la combinaison d'un ou plusieurs scores provenant d'un message, d'une paire de messages ou d'un dialogue (ensemble de messages) (Seo *et al.*, 2009). Les résultats de ces travaux montrent que d'une part la fusion tardive donne de meilleurs résultats que la fusion précoce, et d'autre part que la prise en compte de la structure de la conversation améliore les résultats.

La deuxième catégorie des travaux propose de tenir compte des informations sociales offertes par les structures sociales et d'autres informations intrinsèques telles que la longueur de la conversation. Bhatia et Mitra proposent un modèle probabiliste qui tient compte du contenu textuel des messages, de l'autorité des auteurs, des liens entre conversations et de la taille des conversations (Bhatia et Mitra, 2010). Ces différents facteurs sont pris en compte comme des a priori dans le modèle probabiliste. Les résultats de ces travaux montrent que l'intégration des facteurs sociaux et de la pertinence thématique améliore les résultats par rapport à la pertinence thématique seule.

A notre connaissance, seuls les travaux de Magnani et Montesi proposent de chercher des conversations dans le microblog Twitter (Magnani et Montesi, 2010 ; Magnani *et al.*, 2012). Ces travaux s'inscrivent dans la deuxième catégorie citée plus haut et proposent une approche de fusion des caractéristiques suivantes : i) la pertinence textuelle de la conversation basée sur la fusion tardive de correspondance vectorielle ; ii) la moyenne des tailles des tweets de la conversation ; iii) la popularité de la conversation (nombre de tweet retweetés) ; iv) la popularité des auteurs des tweets calculée par la moyenne des nombres de followers de chaque auteur ; vi) la densité temporelle des tweets. Plusieurs fonctions d'agrégations ont été testées pour calculer la pertinence des conversations : le maximum, le minimum et la moyenne, sur les différents facteurs. Les auteurs montrent que l'utilisation des facteurs sociaux apporte de bons résultats mais il ne montrent pas l'apport de ces facteurs par rapport à une pertinence thématique seule.

Dans cet article, nous proposons un modèle de recherche de conversations dans le microblog Twitter qui tient compte de la pertinence textuelle des tweets et de la pertinence sociale estimée au travers des informations sociales issues du réseau Twitter. Comparativement aux travaux présentés dans l'état de l'art, nous proposons un modèle probabiliste permettant d'intégrer les deux pertinences textuelle et sociale. La pertinence textuelle que nous proposons est estimée par un modèle de langue, à la dif-

férence de Magnani et Montesi qui utilisent un modèle vectoriel. De plus, leur modèle d'ordonnancement des conversations, agrège tous les facteurs en utilisant des formules d'agrégations (maximum, minimum, moyenne) (Magnani et Montesi, 2010 ; Magnani *et al.*, 2012). Nous estimons que ces fonctions d'agrégations ne permettent pas une bonne intégration des ces facteurs. Notre contribution majeure réside dans la définition d'un modèle permettant d'étudier la combinaison de ces facteurs.

3. Modèle de recherche de conversations

3.1. *Présentation informelle des facteurs du modèle*

Nous présentons ici les différents facteurs, autres que textuels, définis pour caractériser une conversation. Le calcul de correspondance entre une requête utilisateur et une conversation va reposer sur le contenu textuel des tweets de cette conversation, mais également sur les facteurs sociaux suivants :

– Participants à la conversation : nous estimons la qualité des participants au travers de leur expertise, leur influence et leur activité. L'expertise est calculée par application d'un modèle de langue inspiré de (Balog *et al.*, 2008 ; BenJabeur *et al.*, 2011). L'influence est calculée par application du *PageRank* sur les relations de retweets renforcées par les relations de favoris (BenJabeur *et al.*, 2011). L'activité d'un utilisateur dans une conversation dépend du nombre de conversations auxquelles il participe et du nombre moyen de tweets du blogueur par conversation ;

– Contenu social : le contenu social des tweets d'une conversation tient compte des métadonnées que sont les urls et les hashtags. Un tweet contenant des urls est plus informatif car il apporte une information portée par l'url (Nagmoti *et al.*, 2010 ; Zhao *et al.*, 2011). Un hashtag catégorise un tweet selon un contexte et augmente la visibilité du tweet (Duan *et al.*, 2010) ;

– Influence sociale d'une conversation : elle est estimée par le nombre des tweets de la conversation qui sont mis en favoris par des blogueurs, ainsi que par le nombre de tweets de la conversation qui sont retweetés.

On remarque que la description ci-dessus utilise certaines informations spécifiques au réseau social Twitter, car ce réseau est celui auquel nous nous intéressons. On peut cependant noter que pour d'autres réseaux sociaux, des informations équivalentes pourraient être soit obtenues directement, soit calculées indirectement.

Une fois cette description générale effectuée, nous pouvons maintenant détailler notre modèle plus formellement.

3.2. *Notations*

Comme nous l'avons décrit précédemment, notre objectif est de rechercher des conversations en tenant compte de leurs tweets et des éléments sociaux qui leur sont

relatifs. Nous définissons les ensembles d'éléments sociaux suivants :

– C : l'ensemble des conversations c considérées ;

– T : l'ensemble des tweets t considérés : l'union de l'ensemble des tweets des conversations et de l'ensemble des tweets des blogueurs que nous utilisons dans le calcul de l'expertise du blogueur (cf section 3.4.1) ;

– TC : l'ensemble des tweets de T apparaissant dans une conversation ;

– B : l'ensemble des blogueurs b considérés : l'union de l'ensemble des blogueurs des conversations et de l'ensemble des blogueurs que nous utilisons dans le calcul de l'influence du blogueur (cf section 3.4.1) ;

– H : l'ensemble des hashtags h (mots-clés) dans la collection C ;

– URL : l'ensemble des urls $urls$ dans la collection C.

Nous définissons les fonctions suivantes :

– $Tweets(b)$: l'ensemble des tweets publiés par le blogueur b ;

– $Dtweets(b)$: la concaténation des tweets publiés par le blogueur b ;

– $Retweets(b)$: l'ensemble des tweets retweetés par le blogueur b ;

– $Favoris(b)$: l'ensemble des tweets mis en favoris par le blogueur b ;

– $Hashtags(t)$: la liste des hashtags présents dans le tweet t ;

– $Urls(t)$: la liste des urls présentes dans le tweet t.

Pour une conversation c , nous notons de plus les fonctions suivantes par une notation pointée pour faciliter la lecture :

– $c.tweet$: l'ensemble des tweets de la conversation c ;

– $c.sujet$: représente le sujet de la conversation c, défini par un vecteur des n termes les plus fréquents dans les tweets de c ;

– $c.blogueur$: l'ensemble des blogueurs de la conversation c ;

– $c.hashtag = \oplus_{t \in c.tweet} Hashtags(t)$, avec $\oplus$ la concaténation de liste de chaînes de caractères ;

– $c.url = \oplus_{t \in c.tweet} Urls(t)$: la liste des urls de la conversation ;

– $c.retweet = c.tweet \cap \left(\bigcup_{b \in B} Retweets(b) \right)$: l'ensemble des tweets de la conversation c retweetés par des blogueurs ;

– $c.favoris = c.tweet \cap \left(\bigcup_{b \in B} Favoris(b) \right)$, l'ensemble des tweets de la conversation c mis en favoris par des blogueurs ;

– $c.nbrRetweets = \sum_{b \in B} |c.tweet \cap Retweets(b)|$, le nombre total de retweets dans la conversation c ;

– $c.nbrFavoris = \sum_{b \in B} |c.tweet \cap Favoris(b)|$, le nombre total de favoris de la conversation c.

3.3. *Description du modèle de correspondance des conversations*

La pertinence d'une conversation c vis à vis d'une requête $Q = \{w_1, w_2, ..., w_m\}$ est estimée en appliquant le théorème de Bayes :

$$P(c|Q) = \frac{P(Q|c)\ P(c)}{P(Q)} \quad [1]$$

La probabilité $P(Q)$ est uniforme pour toutes les conversations c de la collection C, et est considérée comme non discriminante, alors la probabilité de la conversation c pour la requête Q est approximée par :

$$P(c|Q) \propto P(Q|c)\ P(c) \quad [2]$$

La probabilité $P(Q|c)$ est calculée en utilisant un modèle de langue lissé (Jelinek and Mercer 1980) comme suit :

$$P(Q|c) = \prod_{w \in Q} [\lambda_{thematique}\ P(w|c) + (1 - \lambda_{thematique})\ P(w|C)]^{n(w,Q)} \quad [3]$$

Où $P(w|C)$ représente la probabilité d'apparition du terme w dans la collection C, $n(w, Q)$ représente le nombre d'occurrences du terme w dans la requête Q et $\lambda_{thematique}$ est un paramètre de lissage.

Une conversation c est un ensemble de tweets. La probabilité $P(w|c)$ que le terme w apparaisse dans les tweets de la conversation c est donc calculée comme suit :

$$P(w|c) = \sum_{t \in c.tweet} P(w|t)\ P(t|c) \quad [4]$$

Nous considérons que $P(t|c)$ est uniforme pour l'ensemble des tweets et la probabilité $P(w|t)$ représente la probabilité d'apparition du terme w dans le tweet t et est estimée par maximum de vraisemblance.

3.4. *Estimations des probabilités a priori P(c)*

Indépendamment de la requête, nous essayons d'estimer la probabilité d'une conversation. Un tel choix est utile pour permettre de réaliser des recherches rapides avec un maximum d'éléments précalculés. Nous nous basons sur les différents facteurs que nous avons définis en section 3.1.

La probabilité a priori $P(c)$ est estimée à l'aide de trois mesures qui sont : i) la qualité des blogueurs participants à la conversation appelée *BloggersQuality(c)*, ii) le contenu social des tweets de la conversation appelé *SocialContent(c)* et iii) l'influence sociale de la conversation appelée *SocialInfluence(c)*.

3.4.1. *La qualité des participants*

1) **L'activité**

L'activité est un facteur de pertinence qui désigne le degré d'implication et de participation des blogueurs aux conversations. Ainsi une conversation ayant des participants actifs est a priori intéressante. Nous estimons ce facteur comme suit :

$$Activity(b) = \frac{|TC \cap Tweet(b)|}{|TC|} \quad [5]$$

De ce fait, le score d'activité de tous les blogueurs $c.blogueur$ participants à la conversation c est donné par la formule suivante :

$$BloggersActivity(c) = \frac{1}{max_{c' \in C}(|c'.blogueur|)} \sum_{b \in c.blogueur} Activity(b) \quad [6]$$

2) **L'expertise**

Comme le facteur activité, l'expertise des blogueurs participants à la conversation peut être un facteur de pertinence car si ces blogueurs sont experts alors a priori la conversation pourrait être intéressante et de qualité.

Nous considérons l'expertise d'un blogueur b comme un facteur a priori donc indépendant de la requête, ainsi nous proposons d'évaluer cette expertise par rapport au sujet de la conversation $c.sujet$. Ainsi le score d'expertise d'un blogueur pour le sujet de la conversation est donné par la formule suivante :

$$Expertise(b, c) = \prod_{s \in c.sujet} \Big[\lambda_{exp} P(s|Dtweets(b)) + (1 - \lambda_{exp}) P(s|C_{db})\Big]^{n(s, c.sujet)} \quad [7]$$

Où $n(s, c.sujet)$ le nombre d'occurrences du terme s dans le vecteur $c.sujet$, λ_{exp} un paramètre de lissage, $P(s|Dtweet(b))$ est la probabilité d'apparition du terme s dans $Dtweets(b)$ et $P(s|C_{db})$ est la probabilité d'apparition du terme s dans la collection des documents des blogueurs $\bigcup_{b \in c.blogueur} \{Dtweet(b)\}$. Ces probabilités sont calculées par maximum de vraisemblance.

Le score d'expertise d'une conversation c est donné par la formule suivante :

$$BloggersExpertise(c) = \frac{1}{max_{c' \in C}(|c'.blogueur|)} \sum_{\forall b \in c.blogueur} Expertise(b, c) \quad [8]$$

3) **L'influence**

L'influence d'un blogueur est estimée par le nombre de retweets de ses messages dans les conversations et le nombre de ses messages mis en favoris par d'autres blogueurs. Cette influence est encore plus importante si ces tweets sont retweetés ou mis en favoris par des blogueurs eux-mêmes influents. Dans (BenJabeur *et al.*, 2011), les auteurs calculent l'influence d'un blogueur par application du *PageRank* en se basant

sur les relations de rediffusion (retweet). Nous calculons cette influence en intégrant au *PageRank* les relations de retweets et de favoris :

$$Influence(b_i) = d\frac{1}{|B|} + (1-d)\left[\sum_{b_j \in R} w_r(b_i, b_j)\frac{Influence(b_j)}{LR} \times \sum_{b_j \in F} w_f(b_i, b_j)\frac{Influence(b_j)}{LF}\right] \quad [9]$$

Où $d \in [0, 1]$ facteur d'atténuation du *PageRank*, R ensemble des blogueurs ayant partagé une relation de retweet, F ensemble des blogueurs ayant partagé une relation de favoris, $w_f(b_i, b_j)$ poids de la relation de favoris entre b_i et b_j, $w_r(b_i, b_j)$ poids de la relation de retweet entre b_i et b_j, LR le nombre de relations de retweets à partir d'un blogueur b_j vers d'autres blogueurs et LF le nombre de relations de favoris à partir d'un blogueur b_j vers d'autres blogueurs.

Les poids $w_f(b_i, b_j)$ et $w_r(b_i, b_j)$ sont calculés comme suit :

$$w_f(b_i, b_j) = \frac{|Tweets(b_i) \cap Favoris(b_j)|}{|Favoris(b_j)|} \quad [10]$$

Et

$$w_r(b_i, b_j) = \frac{|Tweets(b_i) \cap Retweet(b_j)|}{|Retweets(b_j)|} \quad [11]$$

Le score d'influence de tous les blogueurs participants à la conversation c est donné par la formule suivant :

$$BloggersInfluence(c) = \frac{1}{max_{\forall c \in C}(|c.blogueur|)} \sum_{b \in c.blogueur} Influence(b) \quad [12]$$

Le paramètre de qualité des blogueurs participants à une conversation est combinaison linéaire de l'activité, de l'expertise et de l'influence des blogueurs participants à la conversation c et est donnée par la formule suivante :

$$\begin{aligned} BloggersQuality(c) = {} & \alpha_{Bactivity}.BloggersActivity(c) \\ & + \alpha_{Bexpertise}.BloggersExpertise(c) \\ & + \alpha_{Binfluence}.BloggersInfluence(c) \end{aligned} \quad [13]$$

Où $\alpha_{Bactivity} + \alpha_{Bexpertise} + \alpha_{Binfluence} = 1$ dénotent l'importance relative des facteurs, avec chaque paramètre dans l'intervalle $[0, 1]$.

3.4.2. *Contenu social des tweets*

Mise à part l'information apportée par les termes du contenu textuel du tweet, chaque tweet apporte une information supplémentaire véhiculée par les hashtags et les urls. Ainsi, une conversation comportant ces différents signes (*urls*, *hashtags*) a une probabilité importante d'apporter plus d'informations.

1) **Les hashtags**

Le score hashtags d'une conversation c est estimé par la densité de cette dernière en nombre de hashtags et est donnée par la formule suivante :

$$Hashtags(c) = \frac{|c.hashtag|}{|H|} \quad [14]$$

2) **Les Urls**

Nous considérons qu'une conversation contenant des urls peut a priori apporter plus d'informations. Ce score est donné par la formule suivante :

$$Urls(c) = \frac{|c.url|}{|URL|} \quad [15]$$

Le score *SocialContent(c)* est une combinaison linéaire des deux scores *hashtags(c)* et *urls(c)* et est donnée par la formule suivante :

$$SocialContent(c) = \gamma_{sc}\, Urls(c) + (1 - \gamma_{sc})\, Hashtags(c). \quad [16]$$

Où γ_{sc}, dans l'intervalle $[0, 1]$, est un paramètre du modèle qui dénote l'importance relative des deux critères considérés.

3.4.3. *Influence sociale de la conversation*

Le facteur influence sociale d'une conversation est estimé par le nombre de tweets retweetés et le nombre de tweets mis en favoris. Ce facteur est calculé comme suit :

$$\begin{aligned} P(c) &= \gamma_{sf}\, \frac{c.nbrRetweets}{\sum_{\forall c' \in C} c'.nbrRetweets} \\ &\quad + (1 - \gamma_{sf})\, \frac{c.nbrFavoris}{\sum_{\forall c' \in C} c'.nbrFavoris} \end{aligned} \quad [17]$$

Où γ_{sf} est un paramètre du modèle dans l'intervalle $[0, 1]$.

3.4.4. *Combinaison des facteurs a priori P(c)*

La probabilité a priori de la conversation est une combinaison linéaire des trois facteurs : qualité des blogueurs participant à la conversation, contenu social des tweets de la conversation et influence sociale de la conversation. Elle est estimée comme suit :

$$\begin{aligned} P(c) &= \beta_{Cquality}.BloggersQuality(c) \\ &\quad + \beta_{Ccontent}.SocialContent(c) \\ &\quad + \beta_{Cinfluence}.SocialInfluence(c) \end{aligned} \quad [18]$$

Où $\beta_{Ccontent} + \beta_{Cquality} + \beta_{Cinfluence} = 1$ mesurent l'impact relatif des facteurs, tel que chacun des paramètres est dans l'intervalle $[0, 1]$.

4. La collection de test

Par manque de collection de test spécifique sur les conversations de tweets, nous avons créé notre propre collection en nous basant sur les campagnes TREC Microblog. Une fois la collection de test collectée, nous l'avons indexée avec Lucene [1] en appliquant l'anti-dictionnaire anglais et l'algorithme de lemmatisation de Porter fournis par Lucene. Dans cette section, nous commençons par décrire la collection de test et sa construction, avant de nous intéresser aux évaluations menées et aux résultats obtenus en section 5.

4.1. *Le corpus de conversation*

Pour construire une collection de conversations, nous avons commencé par obtenir un ensemble important de tweets. D'autre part, nous avons décidé d'associer à la collection de conversations proprement dit l'ensemble des blogueurs y participant, avec leurs caractéristiques tirées de Twitter.

4.1.1. *Collecte des tweets*

La collecte des tweets a été établie pour une part en se basant sur des corpus existants pour la recherche d'information :

– Nous avons d'une part utilisé l'API Twitter pour obtenir les tweets correspondant à 165 *topics* des trois corpus de référence pour la recherche de tweets : TREC Microblog 2011, 2012 et 2013 (Ounis *et al.*, 2011), (Soboroff *et al.*, 2012) et (Lin et Efron, 2013). Parmis ces *topics*, seuls 107 *topics* retournent au moins une conversation.

– Comme l'API Twitter ne renvoie que les tweets les plus récents, nous avons utilisé les identifiants apparaissant dans les fichiers *qrels* de ces mêmes campagnes d'évaluation TREC pour retrouver tous les tweets et récupérer l'ensemble des fichiers $statuses/lookup.json$ accessibles par l'API Twitter, qui contiennent l'ensemble des informations relatives aux tweets (id_str, $user\_id$, $in\_reply\_to\_status\_id$, ...).

Pour obtenir des tweets plus récents afin de prendre en compte d'éventuellles évolutions d'usage au cours du temps, nous avons également utilisé l'API Twitter sur des sujets populaires de 2014 et collecté les fichiers $statuses/lookup.json$ des tweets.

4.1.2. *Construction des conversations*

Une fois les tweets collectés, nous avons construit des conversations en s'inspirant des travaux de (Cogan *et al.*, 2012). De manière très succincte, cette étape se base sur l'extraction d'arborescences de tweets d'après les tweets utilisés comme réponses à des tweets.

1. http ://lucene.apache.org/core/

4.1.3. *Collecte des informations des blogueurs*

Afin de pouvoir utiliser les informations sociales nécessaires, nous avons collecté toutes les informations sur les blogueurs (liste des followers, l'ensemble de leurs tweets, liste des favoris, ...).

Le tableau 1 présente les statistiques sur la collection que nous avons construite. Nous constatons que la moyenne de tweets par conversation est de 7. Néanmoins des conversations peuvent atteindre jusqu'à 40 tweets et plus de 10 blogueurs.

Tableau 1. *Statistiques sur les conversations de la collection*

Nombre de conversations	7806
Nombre de tweets	55220
Nombre de blogueurs	10911
Nombre moyen de tweet par conversation	7
Nombre moyen de blogueurs par conversation	5
Nombre moyen de hashtags par conversation	3
Nombre moyen d'urls par conversation	2
Nombre moyen de retweets par conversation	56
Nombre moyen de favoris par conversation	38

4.2. *Requêtes et jugements de pertinence*

1) **Les requêtes** :

Sur la collection de conversations utilisée, nous avons sélectionné des requêtes de deux sources :

- Les requêtes des campagnes TREC Microblog 2011-2012-2013 ;
- Un ensemble de requêtes liées à l'actualité de 2014.

Parmi les 107 requêtes initiales de TREC Microblog 2011-2012-2013, 9 requêtes sont conservées, et 6 requêtes additionnelles sur des sujets d'actualités de 2014 ont été choisies. La construction de la liste des requêtes à sélectionner, sur lesquelles nous évaluons notre modèle s'est basée sur les critères suivants, définis afin d'éviter des biais liés à des caractéristiques en trop petit nombre :

- La requête doit retourner au moins 5 conversations selon l'API Twitter ;
- Les conversations retournées doivent comporter au moins 3 blogueurs ;
- Les conversations retournées doivent contenir au moins 6 tweets.

Cette liste de requête couvre donc des sujets très différents. Le tableau 2 présente les quinze requêtes considérées.

2) **Les jugements de pertinence** :

Les jugements de pertinence sur ces 15 requêtes ont été obtenus comme suit :

Tableau 2. *Les 15 requêtes utilisées pour l'évaluation (en italique celles provenant de TREC Microblog)*

Israel and Turkey reconcile	*Assange Nobel peace nomination*
Obama reaction to Syrian chemical weapons	Gaza-Israel conflict
Bush's dog dies	World Cup scandal
Oprah Winfrey half−sister	Sarkozy's phone tapping
Egypt's Middle Square protest	Crash MH370 Malaysia Airlines
Asteroid hits Russia	Bygmalion Sarkozy affair
Cause of the Super Bowl blackout	Trusted reviews iphone 5s
William and Kate fax save−the−date	

- Nous avons indexé l'ensemble de la collection avec Lucene ;

- Nous avons demandé à 9 utilisateurs (étudiants, entre 25 et 37 ans) de juger (pertinent ou pas) 5 requêtes chacun sur les 100 premières conversations retournées par le système Lucene utilisant BM25 ;

- Chaque conversation est jugée par trois utilisateurs différents et la pertinence finale est celle de la majorité des évaluations.

4.3. *Mesure d'évaluation*

A cause du nombre de requêtes relativement faible, nous utilisons une approche appelée *Leave One Out* (LOO), très utilisée dans le domaine de l'apprentissage automatique (Arlot et Celisse, 2010). Nous l'appliquons à notre cadre de recherche d'information. Considérons un ensemble de N_Q requêtes sur lequel nous voulons évaluer nos systèmes. Une approche LOO consiste à retirer une requête de l'ensemble de requêtes, à optimiser les paramètres du système testé (en terme de MAP) sur les $N_Q - 1$ requêtes restantes, puis à évaluer la requête retirée avec ces paramètres. On réitère cette étape en retirant chacune des requêtes les unes après les autres, de manière à obtenir une valeur d'AP pour chaque requête en ayant optimisé sur les $N_Q - 1$ autres requêtes. La mesure d'évaluation globale utilisée ensuite sur notre corpus d'expérimentation utilise la mesure de Mean Averaged Precision (MAP) sur les AP obtenues par chaque étape du LOO. Le LOO est donc une validation croisée.

Cette évaluation donne donc bien une valeur de MAP, mais elle est optimisée pour chaque sous-ensemble de l'ensemble total de requêtes. Il en ressort que la valeur de MAP est supérieure à une valeur qui serait obtenue en ayant fait une validation à deux plis par exemple, mais l'évaluation que nous proposons permet de contourner la difficulté liée au nombre relativement faible de requêtes utilisées. Dans notre évaluation, nous avons choisis pour l'optimisation de faire une évaluation exhaustive des paramètres du modèle dans l'intervalle $[0, 1]$, par pas de 0.05. Pour BM25, nous avons utilisé les mêmes pas de 0.05, dans les plages de valeurs courantes de ce modèle citées dans (Singhal, 2001).

5. Évaluations expérimentales

Nous menons des expérimentations pour caractériser deux aspects de notre proposition. Tout d'abord, nous comparons notre correspondance thématique avec une approche basée sur le modèle BM25, qui est un bon choix au niveau de la correspondance de contenu, car il est réputé comme l'un des meilleurs modèles actuels. Si notre proposition se comporte conformément aux conclusions de (Seo *et al.*, 2011) (qui stipule qu'une fusion tardive est meilleure qu'une fusion précoce), notre correspondance thématique devrait fournir de meilleurs résultats que le BM25. Les résultats obtenus sont présentés dans le tableau 3. Nous constatons effectivement dans ce tableau que notre proposition apporte une amélioration relative de 7% en terme de MAP. Avec un seuil de significativité statistique de 5% , la différence de MAP obtenue suivant un t-test de Student bilatéral pairé est significative (p=0.001).

Tableau 3. *Résultats de LOO sur la correspondance thématique seule.*

Système	MAP
BM25	0.2846
Notre modèle	0,3048 (+7 %)

Dans un second temps, nous mesurons l'apport des éléments sociaux de notre modèle à la correspondance thématique seule. Nous escomptons une amélioration des résultats avec l'intégration de ces éléments sociaux.

Le tableau 4 présente les résultats obtenus avec l'intégration des facteurs sociaux. Ces résultats montrent que notre intégration des facteurs sociaux se comporte de façon satisfaisante en terme de MAP, avec une amélioration de 65,6%, et avec valeur significative sur la différence en MAP avec le t-test de Student bilatéral pairé (p=0.001).

Tableau 4. *Résultats de LOO sur l'intégration des facteurs sociaux.*

Système	MAP
Notre modèle (thématique)	0.3048
Notre modèle (social)	0,5049 (+ 65,6 %)

Nous tirons les conclusions suivantes de ces résultats :

– la fusion tardive obtient de meilleurs résultats que la fusion précoce ;

– le fait d'utiliser les caractéristiques sociales comme nous l'avons proposé est très utile pour obtenir de bons résultat dans la recherche de conversation.

Nous étudions plus en détails les valeurs de paramètre de notre modèle obtenues durant le LOO. Bien que les valeurs de ces paramètres sont optimisés pour chaque itération du LOO, nous avons cependant constaté une stabilité de ces valeurs optimales.

Pour les paramètres sociaux de la formule [18], le paramètre $\beta_{Cquality}$ de $BloggersQuality(c)$ obtient en moyenne une valeur de 0.6, beaucoup plus impor-

tante que les deux autres paramètres $\beta_{Ccontent}$ de $SocialContent(c)$ et $\beta_{Cinfluence}$ de $SocialInfluence(c)$ qui obtiennent respectivement une valeur moyenne de 0.15 et 0.2. Ceci indique que notre calcul de la qualité des blogueurs participants à une conversation en terme d'expertise, d'influence et d'activité est important pour l'évaluation de la conversation et confirme ainsi notre hypothèse. Pour les paramètres de l'équation [13], nous constatons que le paramètre $\alpha_{Bexpertise}$ de $BloggersExpertise(c)$ obtient une valeur moyenne de 0.7 à la différence des deux autres paramètres $\alpha_{Bactivity}$ de $BloggersActivity(c)$ et $\alpha_{Binfluence}$ de $BloggersInfluence(c)$ qui obtiennent une valeur moyenne de 0.15. Ceci dénote le fait que l'expertise du blogueur par rapport au sujet d'une conversation est très importante.

6. Conclusion

Nous avons présenté dans cet article une modélisation permettant une intégration claire dans un modèle probabiliste des éléments sociaux permettant l'indexation et la recherche de conversations sociales, en particulier sur le réseau social Twitter. Nous avons construit une collection de conversations contenant plus de 50 000 tweets en nous basant sur des données de TREC Microblog, et nous avons obtenu des jugements de pertinence de 9 personnes sur 15 requêtes. Les expérimentations menées montrent que pour, le contenu seul, notre proposition obtient de meilleurs résultats que BM25. D'autre part, l'intégration des éléments sociaux dans notre modèle améliore significativement la qualité des résultats.

Ce travail va être prolongé sur plusieurs directions : i) comme le travail que nous avons décrit ici se focalise en priorité sur les aspects sociaux, nous nous sommes inspiré d'une modélisation du contenu des conversations de l'état de l'art. On peut cependant se poser la question de déterminer quelle représentation probabiliste d'un ensemble de messages est plus appropriée en étudiant des travaux sur les documents structurés, ii) les aspects temporels des messages n'ont pas été inclus dans notre modélisation, nous supposons que cet élément est important et qu'il devra donc être intégré à l'avenir. Les aspects expérimentaux sont également un élément qui devra être étendu pour permettre des évaluations plus fines du comportement des différents paramètres et de leur impact, ainsi qu'une comparaison plus poussée avec l'état de l'art.

Remerciements

Le travail de Nawal Ould-Amer est soutenu financièrement par la Région Rhône-Alpes.

7. Bibliographie

Albaham A., Salim N., « Adapting Voting Techniques for Online Forum Thread Retrieval », *Communications in Computer and Information Science Springer, p 439-448*, 2012.

Arlot S., Celisse A., « A survey of cross-validation procedures for model selection », *The American Statistical Association,the Institute of Mathematical Statistics, p 40-79.*, 2010.

Balog K., Rijke D., Weerkamp W., « Bloggers as experts : feed distillation using expert retrieval models », *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 2008, p 753–754*, 2008.

BenJabeur L., Tamine L., Boughanem M., « Un modèle de recherche d'information sociale dans les microblogs : cas de Twitter », *Conférence sur les Modèles et l'Analyse des Réseaux : Approches Mathématiques et Informatique (MARAMI 2011), Grenoble,*, 2011.

Bhatia S., Mitra P., « Adopting Inference Networks for Online Thread Retrieval », *Association for the Advancement of Artificial Intelligence, Volume 10, p 1300-1305*, 2010.

Cogan P., Andrews M., Bradonjic M., Kennedy W. S., Sala A., Tucci G., « Reconstruction and Analysis of Twitter Conversation Graphs », *The First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research, p 25–31*, 2012.

Duan Y., Jiang L., Qin T., Zhou M., H Y. S., « An empirical study on learning to rank of tweets », *The 23rd International Conference on Computational Linguistics, p 295–303*, 2010.

Elsas J., Carbonell J., « It pays to be picky : an evaluation of thread retrieval in online forums », *The 32rd international ACM SIGIR conference on research and development in information retrieval, p 347–354*, 2009.

Lin J., Efron M., « Overview of the TREC-2013 Microblog Track », *Proceeddings of the 22th Text REtrieval Conference (TREC 2013)*, 2013.

Magnani M., Montesi D., « Toward Conversation Retrieval », *6th Italian Research Conference, IRCDL 2010, Volume 91,p 173-182*, 2010.

Magnani M., Montesi D., Rossi L., « Conversation retrieval for microblogging sites », *Springer Science+Business Medi, Volume 15, Issue 3-4 , p 354-372*, 2012.

Nagmoti R., Teredesai A., DeCock M., « Ranking approaches for microblog search », *Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology IEEE Computer Society, p 153 - 157*, 2010.

Ounis I., Craig M., Jimmy L., Ian S., « Overview of the TREC-2011 Microblog Track », *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

Seo J., Croft B., Smith D., « Online community search using conversational structures », *Springer Information Retrieval Journal, Volume 14, Issue 6 , p 547-571*, 2011.

Seo J., Croft W., Smith D., « Online community search using thread structure », *The 18th ACM Conference on Information and Knowledge Management, p 1907–1910*, 2009.

Singhal A., « Modern Information Retrieval : A Brief Overview », *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, vol. 24, n° 4, p. 35-43, 2001.

Soboroff I., Iadh O., Craig M., Jimmy L., « Overview of the TREC-2012 Microblog Track », *Proceeddings of the 21th Text REtrieval Conference (TREC 2012)*, 2012.

Zhao L., Zeng Y., Zhong N., « A weighted multi-factor algorithm for microblog search », *Proceedings of the 7th international conference on Active media technology , AMT'11,Active Media Technology, Volume 6890, pp 153-161*, 2011.